Proc. Eighth Stat. Sem. 2001, K.U.
ISBN-969-8397-07-8-(123-130)

ROLE OF P-VALUES IN DECISION MAKING

Mubashir-ul-Hassan, Anwer Khurshid and Mohammad
Department of Statistics, University of Karachi
[1]***Department of Mathematics, King Khalid University, Saudi Arabia***

ABSTRACT

The p-value serves a valuable purpose in the evaluation and interpretation of research findings and most computer programs report the results of statistical tests as probability values. It enables the researchers to set their own level of significance and to reject or accept the null hypothesis in accordance with their own criterion rather than that of fixed level of significance. Many statisticians may, in some situations, disagree on its appropriate use and on its interpretation as a measure of evidence. Despite advances in applied statistics, many researchers still rely on a set of tables (i.e. z, t, F, and chi-square) which do not give the upper tail probabilities, for which they have to convert those probabilities to get the required p-values. In recent years statisticians have prepared the tables for z, t, F and chi-square, which gives directly the upper tail probabilities. As many situations, the evaluation of exact p-values for t and F-test become difficult without a computer package. Therefore, direct methods may be used to evaluate the exact p-values. In this article, a description of the p-value along with a review of literature is presented. Also, the tables which provide upper tail probabilities are indicated. Numerical examples are provided and comments and suggestions are made.

1. INTRODUCTION

It is now common practice for statisticians and researchers to report their research findings by a quantity known as the p-value. The p-value is probably the most common statistical index found in the applied sciences literature. The reason for choosing any statistical method for a particular set of data should be that it best serves the objectives of the research that generated the data. There are many situations where computation of p-value is the tool that best serve the objective of the research.

CLASSICAL APPROACH OF HYPOTHESIS TESTING

In classical statistics, the most widely used traditional method of testing any hypothesis is to select critical region and form a rejection rule such that the probability of committing a Type I error does not exceed some preselected number called the level of the test. Then the statistician reports whether or not the observations are "significant" at the chosen level. However in many cases the choice of a significance level is completely arbitrary (see Barnett, 1983, 1991; Casella and Berger, 1990; Fisher 1958).

3. SIGNIFICANCE LEVELS

The significance levels is part of the decision-making approach to statistical inference as discussed, for example, Waterhouse (1979), Barnett (1983), Clarue and Kempson (1997). In the decision making approach to hypothesis testing, the researcher, prior to the collection of the sample data, decides on a level of significance, which is designated by the Greek letter a. A decision rule is established which states in effect, that if the probability, when the null hypothesis is true, of obtaining a value of the test statistics as more extreme than that actually computed from the sample data is equal to or less than preselected level of significance, the null hypothesis is rejected. In other words, a p-value is determined, and if it is equal to or less than the significance level, the null hypothesis is rejected.

4. HYPOTHESIS TESTING STEPS

Testing for statistical significance follows a relatively well-defined pattern, although researchers differ the number and sequence of steps. A general procedure containing several steps, has been developed that may be adapted to a wide variety of hypothesis-testing situations. These steps are given below:

i) State the null and alternative hypotheses H_0 and H_A.
ii) Choose a the level of significance.
iii) Compute the value of test statistic.
iv) Identify the critical region.
v) Make a decision; if sample test statistic falls in the rejection region, reject Ho. Otherwise, do not reject H_0
vi) State the conclusion in words and interpret the result.

Classical approach has some disadvantages:

i) It does not permit researchers having access only to the conclusion of the hypothesis test to make their own evaluation (i.e. to select their own significance level).
ii) It does not provide researchers with the information necessary to assess precisely the strength of the evidence against the null hypothesis

Chow (1996) and Harlow et al. (1997) provide detailed discussion of the rationale of hypothesis testing controversy.

5. THE P-VALUE APPROACH TO HYPOTHESIS TESTING: New Approach

An alternative way to conclude a test of hypotheses is to compare the p-value of the sample test statistic with significance level ((x). The p-value of the sample test statistic is the smallest level of significance for which we can reject Ho. The p-value is compared to significance level α, and on this basis Ho is either rejected or not rejected. Therefore if p - value _< a , we reject Ho, and if p- value > a, we do not reject Ho. Because statistical software packages generate the p-value associated with the sample test statistic and the Ho, the method of using p-values to conclude tests of hypotheses is widely used and very popular. Many research journals require authors to include the p-value of the observed sample statistic. Then readers will have more information and will know the test conclusion for any preset a. The advantage of knowing the p-value is that we know all level of significance for which the observed sample statistic tells us to reject Ho. To alleviate the problems encountered in the classical approach to hypothesis testing, many researches like Bahn (1972), Barnard (1990), Berger and Sellke (1987), and many more, include the p-value of hypothesis test in their work.

A controversy concerning the usefulness of p-value has continued in articles specially in biostatistics and epidemiology with serious proposals offered. Many of the issues that attend p-values can be found in Gibbons and Pratt (1975), Walker (1986), Evans et al. (1988), Rampt and Yancey (1991), Savitz (1993), Shervish (1996), Goodman (1996, 1998), Barnett and Mathisen (1997), Chia (1977) and Huang et al. (1997).

6. R.A FISHER AND THE P-VALUE

R. A Fisher, father of modern statistics, played a major role in bringing the field of biometry and genetics out of their infancies. Fisher proposed the p-value as part of quasi-formal method of inference, which was popularized in his highly influential 1925 book, "Statistical Methods for Research Workers". He was not the first to use the p-value, but he was the first to outline formally the logic behind its use, as well as the means to calculate it in a wide variety of situations (Goodman, 1993). He defined the p-value a significance probability as it is equaled the probability of a given experimental observation, plus more extreme ones, under a null hypothesis. If this number were small, one could "reject" the null hypothesis as unlikely to be true. The use of a threshold p-value as a basis for rejection was called a "significance test".

Many authors in literature have proposed a variety of designations for the p-value. Some of them are listed below:

"Associated probability" by Siegel (1956).
"Critical level" by Lehmann (1959).
"Critical level" or "significance level" by Birnbaum (1962).
"Observed level of significance" by Kraft and van Eeden (1968) and Berry and Lindgren (1996).
"Prob-value" by Wonnacott and Wonnacott (1972).
"Significance probability" by Lehmann (1975).
"Descriptive level of significance" by Kuebler and Smith (1976).

7. INTERPRETATION OF P-VALUE

The computation of a p-value is probably the most common approach for summarizing the results of a statistical test. The p-value, which directly depends on a given sample, attempts to provide a measure of the strength of the result of a test, in contrast to a simple reject or do not reject. If the null hypothesis is true and the chance of random variation is the only reason for sample differences, then the p-value is a quantitative measure to feed into the decision making process as evidence (see Burdette and Gehan, 1970).

The p-value should be interpreted carefully. Although, the p-value has straightforward interpretation it is also misunderstood quantity in statistics. Everitt and Hay (1992) reported that among 70 academic psychologists, only 3 scored 100 percent on a six-item test on the meaning of p. Four usual misinterpretation of the p-values must be emphasized; i.e. p-value is not

i) The probability that null hypothesis is true.
ii) The probability of committing type I error.
iii) The probability of making wrong decision
iv) The probability that the sample statistic is due to chance alone.

Information about the significance level is usually given in the form of the p-value inequality for the most informative a level. For example, if $p < 0.01$, then p is necessarily also less than 0.05 ($p < 0.05$). On the other hand, knowing that $p < 0.05$ does not indicate whether p is also less than 0.01 or not; hence, $p < 0.01$ is more informative than $p < 0.05$ if both are true. Similarly, $p > 0.05$ is more informative than $p > 0.10$ if both are true. Of the a levels for which critical values of the test statistic are available, the most informative level for the particular result at hand is usually reported. Table 1 provides a reasonable interpretation of different p-values.

TABLE-1: Guidelines for using the p-value to access the evidence against the null hypothesis

P-VALUE	INTERPRETATION
p ≤ 0.01	Very strong evidence against H_o or result is highly significant
0.01 5 p < 0.05	Moderate evidence against Ho or result is significant
0.05 <_ p < 0.10	Strong evidence against H_0 or result is marginally significant
p > 0.10	Little or no evidence against H_0 or result is not significant

Once a value of alpha is set, a result either is statistically significant or is not statistically significant. It doesn't matter whether the p-value is very close to alpha or far away. Many scientists and researchers are not ravaged, and refer to result as being "when the p-value is tiny".

Miller (1966) proposed another method of interpreting the p-value. In his conventional interpretation, he proposed terms by contrast, are single, easily distinguished, easily used words with meanings which suggest their idiomatic use.

Motulsky (1995) used a "Micheline Guide" scale to show p-values on graphs. Different investigators use different keys for this scale and one should be sure about the key, as threshold values will be different for different investigators. A general key is shown below in Table 2:

TABLE-: Definitions of proposed terms for statistical' significance

Term	Abbreviation	Conventional Meaning	General Meaning	Michelin Guide Scale	Wording/ Decision
Not significant (ly)	NS	p _> 0.05	p _> a	NS	Not significant
Significant (ly)	SIG	0.05 > p ? 0.01	a > p >_ oa/5	*	Significant
Decisive (ly)	DEC	0.01 > p >_ 0.001	α/5 > p ≥ α/50	**	Highly/very significant
Conclusive (ly)	CON	p < 0.01	p < od50	***	Extremely si nificant

The proposed interpretation of p-value clarify and simplifies the use of statistics in research, and thus encourage investigators to make use of statistics in evaluating their data. Fisher (1958) warned that the p-value was not to be interpreted as a hypothetical frequency of "error" if the experiment were repeated. It was a measure of evidence in a single experiment, to be used to reflect on the credibility of the null hypothesis, in light of the data. P-value is also used as a measure of evidence i.e. p-value is meant to be combined with alternative sources of information about the phenomenon under study. If a threshold for "significance" is used, it was to be flexible and to depend on background knowledge about the phenomenon being studied.

8. BORDER LINE P-VALUES

If the researcher sets the significance level at conventional value of 0.05, then a p-value of 0.049 denotes "statistical significance" and p-value of 0.051 denotes "not statistical significant". Motulsky (1995) argued that it is better to look at the actual p-value, rather than just looking the result is significant or not. That is how we can know whether the p-value is near or far from it (alpha). When a p-value is just slightly greater than alpha, some scientists refer to the result as "marginally significant" or "almost significant". One way to deal with borderline p-value would be to choose between three decisions rather than two. Rather than decide whether a difference is "significant" or "not significant"; middle category of "inconclusion" may be added.

9. BEHAVIOR OF P-VALUE IN HYPOTHESIS TESTING

EXAMPLE 1

Suppose in a setting we have' $H_0 : \mu = 70$ against $H_1 : \mu > 70$ with n = 200, 2 = 74 and a = 30.

The test statistic $Z_{cal} = \underline{x - go}$ -1.89 with critical region $Z > Z_\alpha$ =1.645 is obtained. With z = 74, we obtained Z_{cal} =1.89 and p-value=0.0294. If X = 70 then the test statistic would be z_{cal} = 0 and the p-value = 0.5, which indicates that there is a very little evidence to infer that the population mean is greater than 70. If z = 77 , the test statistic would equal 3.30 with a p-value of 0.0005, indicating that there is a great deal of evidence to infer that the mean exceeds 70.

Several values of z , the resulting test statistics, and p-value have been listed in the following table.

$\bar{x}$	Z = 6∠	P-value
70	0	0.500
71	0.47	0.3192
72	0.94	0.1736
73	1.41	0.0793
74	1.89	0.0294
75	2.36	0.0091
76	2.83	0.0023
77	3.30	0.0005

Above table shows that it is retrieved that as X increases so does the test statistic.

9.1 HYPOTHESIS TEST USING Z-TEST

Example

A farmer claims that the mean weight of turkeys on the farm is 10.2 lb, with a standard deviation of 2. An employee believes the mean weight is more than 10.2 lb. To test this claim, a random sample of 100 turkeys is taken. The sample produces a mean of 10.6. Using the p-value approach, does the sample provide sufficient evidence to reject the null hypothesis at the 0.05 level of significance?

The solution will proceed as follows:

H o : 1t:5 10.2 against H_1 : p > 10.2 with n =100, z =10.6 and a = 2. The test statistic $Z_{cal} = \frac{x - Fro}{\sigma/\sqrt{n}}$ – 2,00. The p-value is found by transforming the test statistic into a probability using table lb, given in the appendix of Hassan (2001), and adjusting the probability to represent the area of the right tail of the curve (if this were a two-tailed test, both tails would be calculated and then added). Table lb gives a probability of 0.4772 for the z score of 2.00 out to the tail, subtract 0.4772 from 0.50. For a one-tailed test, this subtracted value is the p-value which is 0.228 in this example.

Another method of consulting table le, given in the appendix of Hassan (2001), gives directly the upper tail probabilities. See the value of z score 2.00 corresponding to the upper tail probability column, which gives us the value equal to 0.0228 which is also our required p-value.

9.2 HYPOTHESIS TEST USING T-TEST

The calculation for the test statistic is

The transformation of t is slightly different than for z. The z score requires finding a probability in the z table, subtracting it from 0.5000, and adjusting it as necessary. The t score also requires consulting the t table for a probability. To find the probability for t, the degrees of freedom for the sample are also necessary.

Using table 2a, given in the appendix of Hassan (2001), trace down to the appropriate number of degrees of freedom. Transform the calculated value into a probability using table 2a. The p-value for the statistic is determined by moving down the degrees of freedom column, across to the value closest to the calculated value, and up to the alpha associated with the column. The degrees of freedom equal n-1, for a sample of 25, df = 24. The calculated value is 3.04. Move down to df 24. Reading across this row only, notice the value closest to 3.04 is 2.797. Move up the column containing 2.797 and find alpha, 0.005. The alpha value of .005 is the p-value associated with the calculated value 3.04 at 24 degree of freedom.

If the calculated value falls between two critical values, use both alpha values to denote the range of values where the p-value lies. For example, if the calculated value in step 3 equals 2.32 for a sample of 17, with df = 16, then the p-value lies in range from 0.01 to 0.025. Both values are noted and compared to the original alpha in making a decision about the hypothesis.

1 able 2b: Use this table in the same manner as discussed above.

Table 2c: This table is given for the upper tail area for the t distribution. This table can be consult to find out the p-value across the t-statistics. For example if the calculated value is equal to 2.32 (with 16 df), then read this calculated value down the column and across to the degree of freedom, which is equal to p-value 0.017. Hence for the same calculated value, table 2a or table 3b gives the p-value in the range but after consulting table 2c, we have found the exact p-value.

Dawson (1977) presented a novel arrangement of the table that allows p-values to be determined quite precisely from a table of manageable size.

9.3 HYPOTHESIS TEST USING CHI-S-JARE TEST

The calculation for x2 is:

$$\chi^2 = \frac{(n-1)s^2}{62}$$

The transformation of x2 is exactly like t. Table 3c, given in the appendix of Hassan (2001), gives the upper tail areas for x 2 distribution. These have the form $P[\chi^2(n) > c]$ **for the** χ^2 **-tail areas, where n** is the degree of freedom parameter for the corresponding reference distribution.

10. CONCLUDING REMARKS AND SUMMARY

In a report or published article, statistical results often are simply listed, along with the relevant p-value. The authors may not even reach their own conclusions, instead choosing to leave any inferences up to the individual readers. Indeed, many such reportings will not even state the null hypothesis per se, it being impliciffrom the general discussion accompanying the data.

It is recommended that significance level (a)must be chosen before performing the hypothesis test. a reflects the probability level at which experimenter is willing to risk a type I error. Also, the accuracy and reliability of measurement instruments might affect the choice of a. Thus, a should be selected first, then a computer software or a table may be used to find the p-value of test statistic and finally draw the appropriate conclusion.

REFERENCES

1. Balm, A. K. (1972). "P and null hypothesis", *Annals of Internal Medicine,* 76, 674.
2. Barnard, G. A. (1990). "Must Clinical Trails be large? The Interpretation of P-values and the Combination of Test Results", *Statistics in Medicine,* 9, 601-614.
3. **Barnett,** M. L. and Mathisen, A. (1997). "Tyranny of the p-value: The Conflict Between Statistical Significance and Common Sense (Editorial)", *Journal of Dental Research,* **76, 152-154.**
4. **Barnett, V. (1983).** *Comparative Statistical Inference.* John Wiley and Sons, New York.
5. **Barnard, G. A. (1990). "Must Clinical Trails be large? The Interpretation of P-values and the** Combination of Test Results", *Statistics in Medicine,* 9, 601-614.
6. Bennett, J. H. (1991). *Statistical Inference: Selected Correspondence of R. A. Fisher.* John Wiley and Sons, New York.
7. Berger, J. and Sellke, T. (1987). "Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence", *Journal of the American Statistical Association,* 82, 112-139.
8. Berry, D. A. and Lindgren, D. W. (1996). *Statistics: Theory and Methods,* Second edition. Duxbury Press, California.
9. Birnbaum, A. (1962). "On the Foundation of Statistical Inference", *Journal of the American Statistical Association,* 57, 269-306.
10. Burdette, W. J. and Gehan, E. A. (1970). ***Planning and Analysis of Clinical Studies.*** Charles C. Thomas, Springfield, Illinois.
11. Casella, G. and Berger, R. L. (1990). ***Statistical Inference.*** Pacific Grove, Wadsworth, California.

12. Chia, K. S. (1997). "Significant-it is - An Obsession with the p-value", *Scanadavian Journal of Public Health*, 23, 152-154.
13. Chow, S. L. (1996). *Statistical Significance: Rationale, Validity and Utility*. Sage, London.
14. Clarue, G. M. and Kempson, R. E. (1997). *Introduction to the Design and Analysis of Experiments.* Arnold, London.
15. Dawson, R. J. M. (1997). "Turning the tables: A t-table for Today", *Journal of Statistics Education*, (http://www.amstat.org/publications/jse/v5n2/dawson.html).
16. Evans, S. J., Mills, P. and Dawson, J. (1988). "The End of the p-value?," *British Heart Journal*, 60, 177-180.
17. Everitt, B. S. and Hay, D. F. (1992). Talking About Statistics: A Psychologist's Guide to Design and Analysis. Edward Arnold, London.
18. Feinstein, A. R. '(1998). "P-values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin", *Journal of Clinical Epidemiology*, 51, 355-360.
19. Fisher, R. (1958). *Statistical Methods for Research Workers.* 13`h Edition, Hafner, New York.
20. Gibbons, J. D. and Pratt, J. W. (1975). "P-values: Interpretations and Methodology", *American Statistician*, 29, 20-25.
21. Goodman, S. N. (1993). "P-values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate", *American Journal of Epidemiology*, 137,485-495.
22. Goodman, S. N. (1998). "P-values", In: *Encyclopedia of Biostatistics*, (Eds. P. Armitage and T. Colton), Vol. 4, pp. 3233-3237. John Wiley and Sons, New York.
23. Goodman, S. N. (1999). "The P-value Fallacy in Medical Statistics", *Annals of Internal Medicine*, 130, 995-1004.
24. Hassan, M. (2001). "The Role of P-value in Decision Making", *M. Sc. Technical Report.* Department of Statistics, University of Karachi, Pakistan.
25. Huang, H. M. J., O'Neill, R. T., Bauer, P. and Kohne, K. (1997). "The Behavior of the P-value When the Alternative hypothesis Is True", *Biometrics*, 53, 11-22.
26. Kraft, C. H. and van Eeden, C. (1968). A *Nonparametric Introduction to Statistics.* Macmillan, New York.
27. Kuebler, R. R. and Smith, H., Jr. (1976). *Statistics: A Beginning.* John Wiley and Sons, New York.
28. Lehmann, E. L. (1959). *Testing Statistical Hypothesis.* John Wiley and Sons, New York.
29. Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.
30. Miller, D. A. (1966). "Significant and Highly Significant", *Nature*, 210, 1190.
31. Motulsky, H. J. (1995). *Interpreting Nonsignificant p-values.* Intuitive Biostatistics, Oxford University Inc, Oxford.
32. Ramp, W. K. and Yancey, J. M. (1991). "P-values and Their Problems", Bone and Mineral, 13, 163-165.
33. Savitz, D. A. (1993). "Is Statistical Significance Testing Useful in Interpreting Data?", *Reproductive Toxicology*, 7, 95-100.
34. Schervish, M. J. (1996). "P-values: What They Are and What They Are Not", *The American Statistician*, 50, 203-206.
35. Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, New York.
36. Walker, A. M. (1986). "Reporting the Results of Epidemiological Studies", *American Journal of Public Health*, 76, 556-558.
37. Wonnacott, T. H. and R. J. Wonnacott (1972). *Introductory Statistics for Business and Economics.* John Wiley and Sons, New York.